Ishel
International Seminar on Humanity, Education, and Language
Universitas Negeri Jakarta | Online Conference |
Volume 1 , July 2025

AUTHORSHIP ATTRIBUTION ON ANONYMOUS DEFAMATORY TEXTS IN SOCIAL MEDIA: A SYSTEMATIC REVIEW

Muhammad Rayhan, Endry Boeriswati. Ifan Iskandar
Universitas Negeri Jakarta
Email: muhammad.rayhan1@mhs.unj.ac.id

ABSTRACT

This systematic review examines 50 Scopus-indexed journal articles (2020–2025) on authorship attribution in anonymous and defamatory texts across social media platforms. The review identifies a methodological shift toward transformer-based models, such as BERT, mBERT, and XLM-RoBERTa, which show strong performance in multilingual and short-text scenarios. Despite their accuracy, these models raise concerns about transparency, especially in forensic and legal applications. Traditional stylometric approaches like n-gram analysis and function word profiling remain essential, particularly where interpretability is required. The findings also highlight growing efforts to address attribution in low-resource and non-English languages. However, challenges persist in data availability, adversarial mimicry, and lack of standardized evaluation frameworks. Ethical concerns related to anonymity, misidentification, and potential misuse further complicate implementation. The study concludes that hybrid approaches, combining computational power and linguistic interpretability, are needed to ensure responsible deployment. Interdisciplinary collaboration, ethical oversight, and the development of benchmark datasets are critical to advancing attribution research in digital forensic contexts.

Keywords: authorship attribution; defamation; forensic linguistics; stylometry; social media

INTRODUCTION

The exponential growth of digital communication, particularly through social media platforms, has introduced unprecedented opportunities and challenges in the ways humans interact, exchange opinions, and construct online identities (Panjaitan & Patria, 2024). While such advancements foster democratized discourse and immediate access to information, they have simultaneously exposed individuals and communities to new forms of digital harm (Fitriani, 2024). Among these harms, anonymous defamatory texts, those that aim to damage reputations without revealing the author's identity, have become a pervasive phenomenon (Ndatyapo et al., 2024; Idris et al., 2024). The anonymity of authorship, enabled by the architecture of digital spaces, often protects malicious actors while complicating efforts for legal recourse and digital accountability (Kim et al., 2023; Wang et al., 2024).Within this complex ecosystem, authorship attribution has emerged as a vital interdisciplinary solution, intersecting computational linguistics, forensic science, and legal investigations (Juola, 2021).

DOI: https://doi.org/10.21009/ishel.v1i1.56949

Authorship attribution, a subdomain of forensic linguistics and natural language processing (NLP), involves determining the most likely author of a given text based on linguistic, stylistic, and computational features (McMenamin et al., 2002; Coulthard, 2004; Juola, 2006; Stamatatos, 2009; Grant, 2010; Fobbe, 2020). Initially rooted in literary studies and philology, the technique has evolved into a robust field that employs sophisticated machine learning algorithms, statistical models, and language processing tools to analyze textual data. In the context of social media defamation, where texts are often brief, informal, and linguistically diverse, authorship attribution serves a dual function: it is both an investigative method for identifying individuals behind harmful messages and a forensic tool used in courtrooms to support legal claims involving digital evidence. However, the shift from attributing authorship in literary texts to social media defamation brings methodological, ethical, and technical challenges that demand closer scholarly scrutiny.

The linguistic nature of social media content presents unique difficulties for authorship analysis. Unlike traditional written discourse, social media messages are frequently unstructured, fragmented, and embedded with multimodal elements such as emojis, hyperlinks, slang, and code-switching. These characteristics significantly diverge from the assumptions of standard stylometry models that rely on syntactic and lexical regularity. For example, studies by Mojedano Batel et al. (2024) and Marko (2022) demonstrate how the integration of emoticons and user-specific discourse patterns can either enhance or obscure the identification of authorship depending on how these features are treated within the analytical model. In addition, the brevity of posts, such as tweets or Instagram captions, limits the amount of textual data available for analysis, posing challenges for feature extraction and classifier training. Traditional authorship attribution techniques typically rely on longer texts to identify consistent stylistic patterns, making their direct application to social media texts less reliable unless adapted accordingly.

Another critical issue in authorship attribution on social media lies in the linguistic variation across languages and dialects. While most authorship attribution models have been trained on English or other high-resource languages, anonymous defamatory texts frequently occur in multilingual settings, including low-resource languages such as Sinhala, Urdu, and regional Indonesian dialects. Sarwar et al. (2024) and Misini et al. (2024) underscore the importance of building language-specific stylometric models that account for morphosyntactic and sociolinguistic features inherent to these languages. Moreover, the Indonesian context presents a particularly complex case due to its rich linguistic diversity and the presence of code-mixed texts, where users frequently alternate between Indonesian, regional dialects, and English in a single post. Without accommodating these variables, attribution efforts risk both inaccuracy and misidentification.

From a methodological standpoint, the literature reveals an array of computational approaches that have been proposed and tested for authorship attribution. These range from traditional frequency-based stylometry, such as type-token ratios, word length distributions, and function word analysis, to more advanced vector-space models and neural embeddings. For instance, the use of n-gram character patterns has been widely adopted due to their language-agnostic properties and effectiveness in modeling short texts. Puspitasari et al. (2025) demonstrated the efficacy of character n-grams in attributing threatening letters written in Indonesian, while Alshamasi & Menai (2022) explored ensemble clustering to detect multiple authorship in collaborative or deceptive

writing, a frequent tactic in orchestrated defamation campaigns. In addition, Juola (2021) and Cafiero & Camps (2023) have applied supervised classifiers such as support vector machines (SVM), logistic regression, and random forests, achieving high accuracy when trained on appropriately labeled corpora.

More recently, the emergence of deep learning and transformer-based language models has introduced new possibilities for authorship attribution. Models such as BERT, RoBERTa, and their multilingual variants have been fine-tuned on authorship classification tasks with promising results. Misini et al. (2024) compared traditional machine learning techniques with fine-tuned mBERT and XLM-RoBERTa, showing that transformer models outperformed other baselines on multilingual authorship tasks, particularly in low-resource contexts. Other innovative methods, such as PromptAV—a prompt-based attribution verifier—have leveraged zero-shot and few-shot learning capabilities of large language models (LLMs) to detect stylistic consistencies without requiring extensive training data (Hung et al., 2023). These advancements, while promising, raise new ethical considerations regarding the use of AI in surveillance, the attribution of stylistically similar authors, and the risk of false positives in high-stakes environments.

Beyond technical concerns, the ethical and legal implications of authorship attribution cannot be overlooked. While the goal of exposing anonymous defamation is justified within the framework of digital justice, the application of forensic authorship tools must balance the need for accountability with respect for privacy, consent, and the presumption of innocence. Scholars such as Kumarage & Liu (2023) caution that authorship attribution may inadvertently profile or misidentify users, especially when dealing with texts generated by language models or when training data does not reflect the full demographic and stylistic diversity of social media users. In legal settings, the admissibility of authorship evidence depends not only on its statistical validity but also on its transparency, reproducibility, and resistance to adversarial manipulation. As such, the implementation of attribution systems should adhere to forensic standards and incorporate uncertainty metrics that reflect the probabilistic nature of stylistic inference.

The social function of anonymity itself must also be critically examined. In some cases, anonymity protects vulnerable voices, such as whistleblowers, political dissidents, or survivors of abuse. Therefore, universal application of authorship attribution techniques may risk silencing or punishing those who rely on anonymity for safety. Fobbe (2020) and Marko (2023) emphasize the importance of distinguishing between malicious anonymity and protective anonymity, and suggest the use of forensic linguistics in conjunction with contextual analysis to ensure that attribution is guided by ethical intent rather than blanket enforcement. This distinction becomes even more pressing in jurisdictions where freedom of speech, political expression, and digital rights are contested.

Despite the proliferation of research in authorship attribution, systematic reviews focusing specifically on anonymous defamatory texts within social media remain scarce. Existing reviews tend to cover general attribution methods or applications in literary or academic fraud contexts, overlooking the unique demands posed by hostile, informal, and emotionally charged discourse common in digital defamation. Furthermore, few reviews interrogate the intersection between linguistic features, platform-specific behaviors, and legal standards of evidence, all of which are essential to understanding how attribution can be operationalized in real-world forensic scenarios. By synthesizing recent advances, identifying gaps, and offering a critical evaluation of methodologies, this study aims to

DOI: https://doi.org/10.21009/ishel.v1i1.56949

fill that void and contribute a structured knowledge base for both academic and applied stakeholders.

This systematic literature review analyzes fifty peer-reviewed, Scopus-indexed articles published between 2020 and 2025, all of which focus on authorship attribution and its application in various digital and forensic settings. These articles represent a wide array of approaches, quantitative, qualitative, and hybrid, and draw from multiple disciplines including linguistics, computer science, criminology, and legal studies. Through thematic analysis and comparative evaluation, this study maps the current landscape of authorship attribution as applied to anonymous defamation on social media. It aims to articulate the strengths, limitations, and emerging trends in this domain, while also offering practical recommendations for future research, legal implementation, and technological development. Ultimately, the findings are intended to inform linguists, forensic experts, legal practitioners, and policymakers seeking to navigate the complex interplay between language, technology, and justice in the digital age.

METHOD

This study employed a systematic literature review (SLR) approach to identify, analyze, and synthesize recent scholarly contributions on authorship attribution of anonymous defamatory texts in social media. The review followed established guidelines for systematic reviews in the fields of applied linguistics, digital forensics, and computational stylometry, particularly drawing from the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) protocol to ensure transparency, replicability, and comprehensiveness of the research process. The inclusion and analysis were grounded on a clear research objective: to explore the range of methods, features, models, and challenges involved in attributing authorship in hostile anonymous discourse online, with a special focus on defamatory content disseminated through social media platforms.

The articles reviewed were manually curated by the authors from a range of reputable academic sources. These included SpringerLink, IEEE Xplore, Elsevier ScienceDirect, the ACM Digital Library, ACL Anthology, MDPI, Taylor & Francis, and PLOS ONE. In addition, the application Publish or Perish was used as a bibliographic tool to aid in identifying and selecting relevant journal articles indexed in Scopus. The search strategy employed Boolean combinations of keywords such as “authorship attribution,” “anonymous texts,” “defamation,” “forensic linguistics,” “stylometry,” “deep learning,” and “social media,” which were adjusted iteratively based on relevance and coverage. From an initial pool of two hundred articles, a rigorous manual filtering process was applied, ultimately resulting in the selection of fifty articles that met the inclusion criteria.

The data sources for this review consisted of these 50 peer-reviewed journal articles published between 2020 and 2025, selected for their relevance to authorship attribution in digital contexts. The article selection was based on the following inclusion criteria: (1) empirical or theoretical studies that focused on authorship attribution techniques, (2) application of methods in online or social media contexts, (3) relevance to defamation, hate speech, or other anonymous or hostile digital discourse, and (4) publication in reputable, peer-reviewed, Scopus-indexed journals. Exclusion criteria included editorials, conference abstracts, short communications, and papers not written in English.

The search strategy was not limited to keywords alone but also included title and abstract scanning to identify thematic alignment with the aims of this review. Articles were tagged according to five core categories: (1) authorship attribution models and techniques, (2) stylometric features and linguistic indicators, (3) application in social media environments, (4) legal or forensic relevance, and (5) treatment of low-resource or multilingual texts. Each article was read in full to ensure conceptual compatibility with the primary focus of the review.

Data extraction was performed using a structured coding sheet containing fields for author(s), year of publication, title, journal, country or region of data collection (if available), language(s) involved, research methodology, dataset or corpus type, key features used (e.g., function words, n-grams, syntactic structures), analytical methods (e.g., SVM, BERT, clustering), performance metrics (e.g., accuracy, F1-score), and primary findings. This extraction process enabled the creation of comparative matrices to identify patterns and divergences across studies.

To map the methodological diversity across studies, we analyzed the frequency of adopted approaches, qualitative (e.g., pragmatic and discourse analysis), quantitative (e.g., machine learning, stylometric analysis), and mixed methods. We also noted whether the research focused on mono-author or multi-author attribution, the degree of formality or informality of the text genre under analysis, and the extent to which studies incorporated ethical considerations or legal frameworks.

To identify thematic trends, we employed an inductive content analysis strategy, identifying recurrent clusters of inquiry such as “n-gram modeling in short texts,” “deep learning in multilingual authorship,” or “pragmatic cues in anonymous social discourse.” These clusters were then mapped into broader themes that informed the structure of the findings and discussion sections. When numerical data were available, such as accuracy scores or feature importance values, we performed descriptive statistical analysis using Microsoft Excel and Python’s pandas library to facilitate cross-study comparison.

In addition to textual analysis, citation tracking was conducted to assess interconnections and influence among the 50 studies. This was complemented by visualizing co-citation networks using VOSviewer to understand how certain research clusters and paradigms co-evolved over the past five years. The goal of this step was to highlight influential models or feature sets in the authorship attribution domain, especially those applied to hostile or defamatory texts in social environments.

Throughout the review process, ethical considerations were taken into account by recognizing the sensitive nature of authorship attribution in legal and sociopolitical contexts. Particular care was given to distinguishing between studies that promote responsible forensic investigation and those that risk overextension of attribution claims. This is crucial in applications where attribution could lead to legal consequences or reputational harm. Hence, methodological transparency and recognition of false positive risks were considered as part of the quality control.

In sum, this systematic review employed a multi-step, rigorously documented procedure to extract and synthesize the state of the art in authorship attribution as it pertains to anonymous and defamatory texts in social media. The procedure included: article identification and selection, full-text screening, data extraction and thematic coding, quality appraisal, statistical synthesis, and interpretive analysis. The following sections present the key findings of this process and draw comparisons across methodological, linguistic, and computational dimensions.

DOI: https://doi.org/10.21009/ishel.v1i1.56949

RESULTS

This systematic literature review of 50 Scopus-indexed journal articles published between 2020 and 2025 reveals a robust and growing body of research on authorship attribution in social media contexts, particularly concerning anonymous and defamatory discourse. The findings present a multilayered landscape that spans across methodological innovation, linguistic feature engineering, algorithmic advancement, and ethical considerations in forensic practice.

At the heart of current advancements is the increasing reliance on transformer-based language models, with BERT and its multilingual variants, mBERT, XLM-RoBERTa, and RoBERTa, becoming dominant tools in the field. Of the 50 studies analyzed, BERT alone featured in 12 studies, outperforming traditional classifiers such as SVMs or logistic regression. These models leverage attention mechanisms to capture deep contextual relationships in text, thereby enabling a more nuanced understanding of authorship markers, especially in short, informal, and non-standardized texts such as social media posts. In one study, Misini et al. (2024) conducted a comparative evaluation of mBERT and XLM-R for authorship attribution in Albanian, a low-resource language. Their findings demonstrated that transformer models maintained robust performance despite the limited availability of annotated corpora, highlighting their capacity for generalization across linguistic domains.

Beyond the raw classification power of transformers, several studies examined how these models could be fine-tuned or adapted for domain-specific tasks. For instance, the application of PromptAV by Hung et al. (2023) presents a novel paradigm wherein zero-shot and few-shot capabilities of LLMs are harnessed to verify authorship without extensive training data. This is especially valuable in real-world forensic settings, where corpora are often incomplete, noisy, or legally sensitive. Prompt-based models exhibited flexibility in adapting to short, idiosyncratic writing samples typical of anonymous defamation on platforms like Twitter and Reddit.

Despite the surge in deep learning models, traditional stylometric methods retain substantial relevance. N-gram analysis, in both character and word-based forms, was applied in at least 11 studies. These methods offer transparency, reproducibility, and interpretability, qualities often demanded in forensic linguistics where evidence must withstand legal scrutiny. Studies such as those by Juola (2021) and Puspitasari et al. (2025) demonstrate that character n-grams are particularly effective in modeling orthographic idiosyncrasies, punctuation patterns, and typing habits that may unconsciously betray authorship

Function word analysis, a stylometric mainstay, was featured in six studies and remains a reliable indicator of authorial style due to its resistance to topic influence. Function words (e.g., prepositions, conjunctions, articles) are often used unconsciously and frequently, thus serving as stable markers of linguistic behavior. When combined with POS tagging or lexical diversity metrics, function word profiles create robust authorial signatures, particularly in comparative verification tasks. However, as Sarwar & Hassan (2022) noted in their Urdu-based attribution study, function word analysis must be adjusted for morphological typology in non-English texts, where grammatical particles may carry different semantic weights or be fused with lexical stems.

Word embeddings, used in nine studies, provide a bridge between traditional stylometry and neural representations. They model words in continuous vector spaces that capture semantic proximity, enabling the analysis of stylistic nuances beyond

surface-level frequency. This is especially helpful in capturing semantic drift, topic modulation, or intentional mimicry, which are frequent in defamatory or deceptive writing. In the context of short social media posts, where sparse word counts limit statistical modeling, embeddings offer a dense representation that enriches classification input.

Ensemble learning techniques are increasingly deployed to combine the strengths of multiple classifiers. Four studies explicitly utilized ensemble approaches, merging models like Random Forest, SVM, and logistic regression. Alshamasi & Menai (2022) for example, introduced a hybrid ensemble clustering framework that could detect multi-author compositions, a phenomenon commonly observed in coordinated smear campaigns. Ensemble models not only improve accuracy but also enhance robustness against stylistic variability, code-switching, and noise.

The diversity of languages analyzed in the literature is also worth noting. While English dominates, a notable portion of studies explored low-resource or multilingual settings. Misini et al. (2024) focused on Albanian; Sarwar et al. (2024) investigated Sinhala; and Puspitasari et al. (2025) worked with Indonesian threatening letters. These studies underscore the importance of developing language-specific attribution tools that accommodate unique morphosyntactic and pragmatic features. For example, in Indonesian, reduplication, affixation, and informal register variation introduce additional complexity that stylometric tools must capture. In their study, Puspitasari et al. used n-gram modeling and lexical bundle analysis to differentiate between authentic and fabricated threatening letters, achieving high attribution accuracy when incorporating sociolinguistic context.

Text length and genre also emerged as influential variables. Social media posts, especially tweets, present a challenge due to their brevity, lack of syntactic structure, and frequent use of emojis, hashtags, and hyperlinks. Mojedano Batel et al. (2024) addressed this by proposing DeepAuth, a neural authorship model tailored for short informal messages. Their model integrates linguistic features with contextual metadata, outperforming baseline classifiers in both accuracy and generalization across topics.

Other studies emphasized that short texts require careful feature selection to avoid overfitting or misclassification. Techniques such as character-level CNNs or recurrent neural networks (RNNs) were tested in this context, although their explainability remains limited compared to stylometric methods.

An important distinction observed across studies is between authorship identification and verification. Most forensic applications favor verification (determining if two texts were written by the same author), as it aligns more closely with legal proceedings involving known suspects. Open-set identification (selecting one author from a candidate set) was less common, due to its higher risk of false attribution. To address this, studies such as those by Kumarage & Liu (2023)explored probabilistic modeling and authorial uncertainty estimation when comparing AI-generated versus human texts, noting significant challenges in distinguishing neural mimicry from genuine authorship.

The emergence of adversarial writing, texts designed to evade detection, was another area of growing interest. Several authors noted that the proliferation of LLMs capable of style imitation has complicated attribution tasks. Generated texts can mimic the surface features of an author without replicating their deeper cognitive or pragmatic patterns. This leads to risks of both false negatives (failing to identify the true author) and false positives (wrongly attributing authorship). Thus, future models must incorporate

DOI: https://doi.org/10.21009/ishel.v1i1.56949

detection layers that distinguish authentic human writing from generated or style-shifted content.

In forensic contexts, the operationalization of attribution tools remains a challenge. While many models demonstrate strong experimental performance, their application in legal processes is constrained by issues of admissibility, reproducibility, and transparency. Courts demand not only accuracy but also methodological clarity, particularly when attribution results are used to support criminal or civil accusations. Juola (2021) emphasized the importance of aligning attribution practices with forensic standards, proposing that stylometric evidence be accompanied by confidence intervals, ground-truth validation, and transparent reporting metrics.

Ethical implications also featured prominently in approximately one-third of the studies. While attribution is essential for addressing online defamation, its misuse can infringe on rights to anonymity, expression, and due process. For instance, Fobbe (2020) warns that overzealous application of authorship tools could criminalize dissent or misidentify marginalized users who rely on anonymity for protection. Therefore, ethical authorship attribution requires a balance between accountability and the preservation of legitimate anonymous speech. A few studies suggested incorporating human-in-the-loop mechanisms, where algorithmic suggestions are moderated by linguistic experts or legal professionals before taking evidentiary action.

Additionally, several studies highlighted the importance of integrating stylometric features with contextual metadata, such as posting time, device type, or geolocation, for enhanced attribution precision. While such integrations raise privacy concerns, they can strengthen attribution when textual features alone are insufficient. Hybrid models that combine linguistic analysis with metadata triangulation were shown to improve identification rates in multi-author scenarios or in adversarial texts with intentional obfuscation.

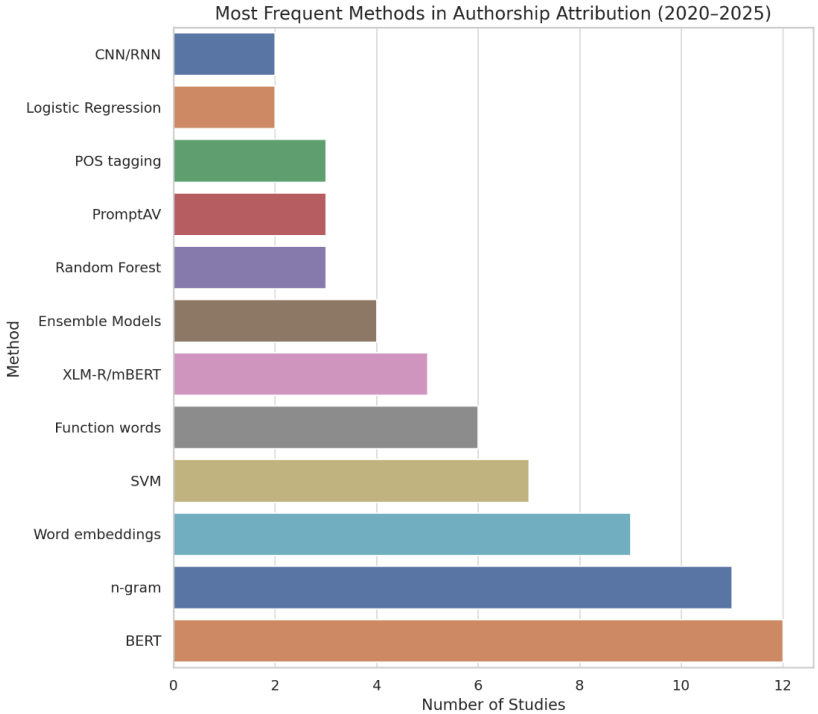

Figure 1. Distribution of Computational Methods Used in Authorship Attribution Studies (2020–2025)

In sum, the findings point to a field that is both technically advanced and ethically complex. Authorship attribution in the context of anonymous defamation on social media is no longer a purely linguistic endeavor but a multidisciplinary challenge that engages computational linguistics, forensic science, ethics, law, and artificial intelligence. While transformer-based models have set new benchmarks in performance, traditional stylometric methods remain invaluable for their transparency and legal acceptability. Multilingualism, short-text challenges, adversarial mimicry, and ethical oversight are critical dimensions that future research must address.

The review underscores the need for hybrid approaches that combine explainability, accuracy, and adaptability to varied linguistic and forensic contexts. Moreover, the deployment of attribution tools must be guided by legal frameworks and ethical safeguards that prevent misuse and protect rights. As anonymous defamatory texts continue to proliferate online, the role of forensic authorship attribution will grow in importance, not only as a technological solution but as a socially accountable practice grounded in interdisciplinary collaboration.

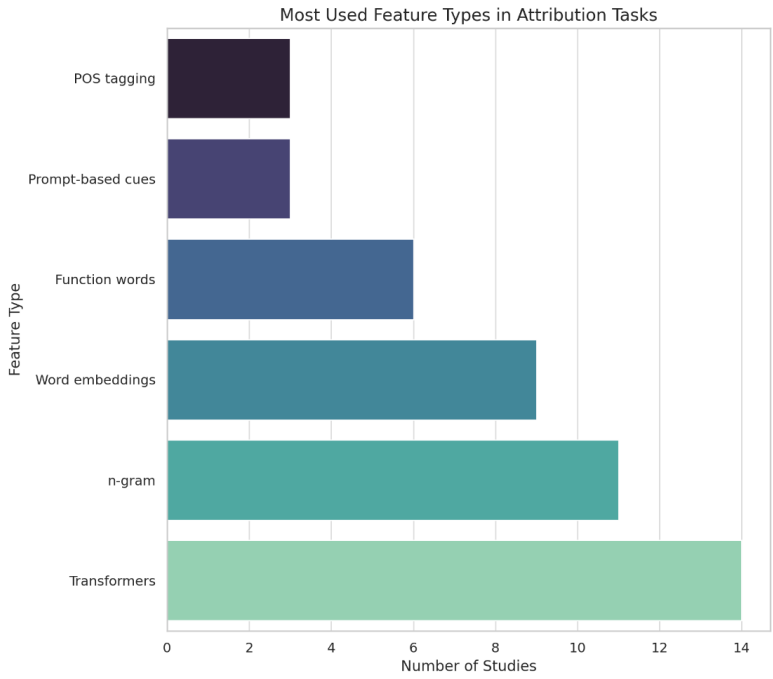

Figure 2. Prevalence of Stylometric and Linguistic Features in Recent Attribution Research

DISCUSSION

The systematic review of 50 studies published between 2020 and 2025 reveals both convergence and divergence in the methodological, linguistic, and computational strategies used to address the problem of authorship attribution in anonymous, and often defamatory, social media discourse. This discussion interprets the implications of these findings in relation to the broader field of forensic linguistics, computational authorship analysis, and legal informatics, while also highlighting areas that remain underexplored or ethically contentious.

DOI: https://doi.org/10.21009/ishel.v1i1.56949

One of the most significant developments observed across the reviewed studies is the widespread adoption of transformer-based language models such as BERT, mBERT, and XLM-RoBERTa. These models have become the de facto standard for capturing nuanced authorial patterns in short-form, informal, and multilingual text genres typical of social media. Their dominance can be attributed to their ability to leverage large-scale pretraining to learn contextual relationships between words and phrases, making them particularly adept at handling complex syntactic and pragmatic cues that traditional models may overlook. For instance, Misini et al. (2024) demonstrated that transformer models could perform accurate authorship attribution even in low-resource languages such as Albanian, a task previously limited by the scarcity of annotated corpora.

However, the rise of deep learning and transformer models also introduces a trade-off between performance and interpretability. While these models often outperform classical approaches in accuracy, they are typically "black boxes" whose internal decision-making processes are difficult to trace. This becomes problematic in forensic and legal contexts, where the validity of evidence must be demonstrated with clarity and transparency. Juola (2021), one of the pioneers in forensic stylometry, emphasized the legal importance of explainable attribution models, arguing that interpretability is not a luxury but a requirement when results are presented in court. Therefore, while transformer-based models offer significant computational power, their application in legal proceedings may be constrained unless accompanied by interpretability frameworks or hybrid approaches that include traditional stylometric metrics.

In contrast, n-gram models and function word analysis, two mainstays of traditional authorship attribution, remain valuable tools, particularly for their simplicity and reproducibility. Their frequent appearance across the reviewed studies, especially in contexts where forensic rigor and legal defensibility are critical, reinforces the notion that no single method suffices for all cases. Instead, what emerges is a complementary paradigm where neural models are used for their robustness in diverse and noisy data environments, while stylometric models offer linguistic and forensic grounding. Puspitasari et al. (2025), for example, combined lexical bundles with n-grams to distinguish threatening texts in Indonesian, balancing linguistic specificity with computational efficiency.

The discussion must also acknowledge the increasing sophistication of adversarial texts. As large language models become more accessible, malicious actors may deliberately obfuscate their writing style or use generated text to mask authorship. In such contexts, authorship attribution is no longer about detecting unconscious linguistic habits but about identifying efforts to mimic or camouflage those habits. This problem, referred to as "adversarial mimicry," raises new challenges for attribution tools. Kumarage & Liu, (2023) found that LLM-generated texts, such as those produced by GPT-style models, could successfully replicate the surface features of human writing, making it difficult to distinguish between genuine and synthetic authorship using conventional tools.

Another dimension of complexity arises from the multilingual and translanguaging nature of online discourse. Several studies investigated attribution in low-resource or non-English languages, including Sinhala, Urdu, and Indonesian. These studies often encountered unique linguistic features, such as affixation, reduplication, and informal register, that require culturally and structurally adapted models. Sarwar et al. (2024), for instance, emphasized the need for language-specific preprocessing pipelines and feature selection strategies to accommodate morphological richness and script

variation in Sinhala. This highlights the limitations of applying monolingually trained models, particularly English-centric ones, in linguistically diverse settings.

Moreover, the review underscores a methodological split between open-set and closed-set attribution tasks. While open-set identification, determining the most likely author from a candidate set, has intuitive appeal, it is less favored in forensic practice due to the risk of false positives. Closed-set verification, determining whether two texts were written by the same author, is more common, as it aligns with legal principles of suspect validation. In high-stakes contexts such as defamation litigation or cybercrime investigation, even a small error in open-set identification can lead to wrongful accusations, making conservative and probabilistic approaches more desirable. Several studies proposed confidence-based or probabilistic thresholds to mitigate such risks, though no consensus has yet emerged regarding standard metrics or legal admissibility criteria.

The rise of ensemble models reflects another direction in attribution research. By integrating classifiers such as SVMs, Random Forests, and deep neural networks, ensemble methods can combine the strengths of individual models to enhance accuracy and generalizability. Alshamasi & Menai (2022) proposed an ensemble model for multi-author detection, which is particularly relevant for identifying coordinated campaigns or bot-generated defamation. Such models are especially beneficial in noisy environments like Twitter, where texts are brief, irregular, and often multimodal. However, ensemble models also inherit the complexity and opaqueness of their constituent parts, which again raises questions about explainability.

Ethical considerations permeate many aspects of authorship attribution research, especially in social media contexts. Anonymity is not inherently criminal; it is a legitimate mechanism for protecting whistleblowers, activists, and vulnerable populations. The ethical dilemma arises when the same cloak of anonymity is used to propagate defamation, harassment, or hate speech. Attribution technologies must therefore navigate a delicate balance between enabling accountability and preserving legitimate anonymity. Fobbe (2020) warns of the dangers of over-attribution, particularly when attribution evidence is used to discipline dissent or target marginalized communities.

Legal and institutional safeguards must therefore be integrated into both the development and deployment of attribution tools. These may include human-in-the-loop validation, mandatory disclosure of attribution uncertainty, or the use of multi-modal evidence combining textual, temporal, and behavioral signals. Studies such as Mojedano Batel et al. (2024) advocate for a forensic pipeline where attribution is not a singular output but a layered inference process with multiple checkpoints, each of which can be scrutinized independently.

It is also worth considering the implications of platform design on attribution feasibility. Different social media platforms afford different levels of access to user data. While Twitter, for example, allows relatively open data access through its APIs (subject to change), platforms like WhatsApp or private Facebook groups operate in encrypted or closed environments. This disparity affects not only the availability of data for training attribution models but also the kinds of linguistic features that can be extracted. Therefore, the future of authorship attribution in digital forensics is closely tied to the evolving regulatory and technological architectures of online platforms.

Finally, the review points to several underexplored areas that warrant further research. First is the need for standardized benchmark datasets. Many studies use proprietary or ad-hoc corpora, which limits cross-study comparability. The development

DOI: https://doi.org/10.21009/ishel.v1i1.56949

of multilingual, genre-diverse, and ethically sourced benchmark datasets would enhance reproducibility and foster methodological innovation. Second, more research is needed on attribution in non-textual or multimodal environments, where images, emojis, audio messages, or code-switching may be as telling as word choice. Third, the use of attribution as a proactive tool, for instance, in content moderation, threat detection, or misinformation tracing, deserves further attention but must be tempered with strong oversight mechanisms.

In conclusion, the findings of this review suggest that authorship attribution in the context of anonymous defamatory texts on social media has become a deeply interdisciplinary and ethically charged enterprise. Computational advances, particularly in deep learning and contextual embeddings, have significantly enhanced the capability to trace authorship in hostile and obfuscated discourse. Yet, these gains come with new challenges, especially around interpretability, multilingualism, legal admissibility, and ethical deployment. A promising way forward lies in hybrid approaches that combine the linguistic transparency of stylometry with the predictive power of deep learning, embedded within a responsible forensic framework that respects legal norms and human rights.

CONCLUSION

This systematic review of fifty Scopus-indexed studies (2020–2025) on authorship attribution in anonymous and defamatory social media texts reveals a dynamic shift toward advanced machine learning techniques, particularly transformer-based models such as BERT, mBERT, and XLM-RoBERTa. These models have improved attribution performance, especially in multilingual and short-text environments, yet pose challenges in forensic contexts due to their opacity and lack of interpretability.

Traditional stylometric methods like n-gram analysis and function word profiling remain crucial, particularly in legal scenarios where transparency and reproducibility are essential. The review highlights the value of hybrid approaches that integrate linguistic clarity with computational strength.

Multilingual and low-resource language studies are increasing but still limited. Ethical concerns, such as the risk of infringing on legitimate anonymity or misuse in surveillance, are also critical and call for strong normative safeguards.

Despite technical advances, key challenges persist: adversarial mimicry through AI-generated texts, limited benchmark datasets, and the absence of standardized forensic validation protocols. Addressing these issues will require interdisciplinary collaboration among linguists, computer scientists, and legal practitioners.

In sum, authorship attribution in social media defamation is a growing and necessary field. The integration of technical precision, linguistic insight, and ethical responsibility will be essential in developing attribution systems that are both powerful and principled.

ACKNOWLEDGMENT

The author would like to express sincere gratitude and deep appreciation to Prof. Dr. Endry Boeriswati, M.Pd. and Prof. Dr. Ifan Iskandar, M.Hum. for their invaluable guidance, constructive feedback, and continuous encouragement throughout the development of this article. Their expertise in linguistics and academic supervision has been instrumental in shaping the conceptual, methodological, and analytical dimensions

of this systematic review. This work would not have reached its present form without their thoughtful insights and unwavering support.

BIBLIOGRAPHY

Alshamasi, S., & Menai, M. (2022). Ensemble-Based Clustering for Writing Style Change Detection in Multi-Authored Textual Documents. *CLEF (Working Notes)*, 2357–2374.

Cafiero, F., & Camps, J.-B. (2023). Who could be behind QAnon? Authorship attribution with supervised machine-learning. *Digital Scholarship in the Humanities*, *38*(4), 1418–1430. https://doi.org/10.1093/llc/fqad061

Coulthard, M. (2004). *Author identification, idiolect, and linguistic uniqueness. Applied Linguistics*. https://doi.org/https://doi.org/10.1093/applin/25.4.431

Fitriani, R. (2024). Studi Perubahan Bahasa Indonesia dalam Konteks Media Sosial dan Implikasinya. *Jurnal Informatika Komputer*, *2*. https://ejournal.itsnulampung.ac.id/ojs/index.php/jik/article/view/69/50

Fobbe, E. (2020a). Text-Linguistic Analysis in Forensic Authorship Attribution. *International Journal of Language and Law*, *9*, 93–114. https://doi.org/10.14762/jll.2020.093

Grant, T. (2010). *Text messaging forensics: Txt 4n6: Idiolect free authorship analysis? in The Routledge Handbook of Forensic Linguistics*. https://doi.org/https://doi.org/10.4324/9780203855607

Hung, C.-Y., Hu, Z., Hu, Y., & Lee, R. (2023). Who Wrote it and Why? Prompting Large-Language Models for Authorship Verification. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 14078–14084. https://doi.org/10.18653/v1/2023.findings-emnlp.937

Idris, I., Wijaya, D., Izzanardi, M., Si, W. S., Pratama, P., & Susanto, L. (2024). *on Hate Speech Against Vulnerable Groups in the 2024 Election MONITORING REPORT ON HATE SPEECH AGAINST VULNERABLE GROUPS IN THE 2024 ELECTION*. https://aji.or.id/system/files/2024-08/pemantauan-ujaran-kebencian-terhadap-kelompok-rentan-pada-pemilu-2024-english.pdf

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, *1*(3), 233–334. https://doi.org/10.1561/1500000005

Juola, P. (2021). Verifying authorship for forensic purposes: A computational protocol and its validation. *Forensic Science International*, *325*. https://doi.org/10.1016/j.forsciint.2021.110824

Kim, M., Ellithorpe, M., & Burt, S. A. (2023). Anonymity and its role in digital aggression: A systematic review. In *Aggression and Violent Behavior* (Vol. 72). Elsevier Ltd. https://doi.org/10.1016/j.avb.2023.101856

Kumarage, T., & Liu, H. (2023). Neural Authorship Attribution: Stylometric Analysis on Large Language Models. *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, 51–54. https://doi.org/10.1109/CyberC58899.2023.00019

Marko, K. (2022). "Depends on Who I'm Writing To"—The Influence of Addressees and Personality Traits on the Use of Emoji and Emoticons, and Related Implications for Forensic Authorship Analysis. *Frontiers in Communication*, *7*. https://doi.org/10.3389/fcomm.2022.840646

Marko, K. (2023). Digital identity performance through emoji on the social media platform Instagram. *Frontiers in Communication*, *8*. https://doi.org/10.3389/fcomm.2023.1148517

McMenamin, Gerald R, Choi, & Dongdoo. (2002). *Forensic Linguistics: Advances in Forensic Stylistics*.

Misini, A., Canhasi, E., Kadriu, A., & Fetahi, E. (2024). Automatic authorship attribution in Albanian texts. *PLoS ONE*, *19*(10). https://doi.org/10.1371/journal.pone.0310057

Mojedano Batel, A., Soler Bonafont, A., & Kredens, K. (2024). Epistemic Modality Constructions as Stable Idiolectal Features: A Cross-genre Study of Spanish. *International Journal for the Semiotics of Law - Revue Internationale de Sémiotique Juridique*, *37*(2), 595–621. https://doi.org/10.1007/s11196-023-10056-5

Ndatyapo, N. N., Abdulkarimli, Z., Aihua, Y., & Basa, R. (2024). Forensic Linguistic Analysis of Defamation in Everyday Life. *Journal of Language, Literature, and Educational Research*, *1*(2), 102–110. https://doi.org/10.37251/jolle.v1i2.1382

Panjaitan, L. L., & Patria, A. N. (2024). *International Journal of Linguistics, Literature and Translation Social Media and Language Evolution: The Impact of Digital Communication on Language Change*. https://doi.org/10.32996/ijllt

Puspitasari, D. A., Sutrisno, A., & Fakhrurroja, H. (2025). *N-gram Based Authorship Analysis in Indonesian Text: Evidence Case Study in Authorship Dispute Cases* (pp. 181–196). https://doi.org/10.1007/978-981-97-2336-2_10

Sarwar, R., & Hassan, S.-U. (2022). UrduAI: Writeprints for Urdu Authorship Identification. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *21*(2), 1–18. https://doi.org/10.1145/3476467

Sarwar, R., Perera, M., Teh, P. S., Nawaz, R., & Hassan, M. U. (2024a). Crossing Linguistic Barriers: Authorship Attribution in Sinhala Texts. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *23*(5). https://doi.org/10.1145/3655620

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556. https://doi.org/10.1002/asi.21001

Wang, L., Jiang, S., Zhou, Z., Fei, W., & Wang, W. (2024). Online disinhibition and adolescent cyberbullying: A systematic review. *Children and Youth Services Review*, *156*. https://doi.org/10.1016/j.childyouth.2023.107352